Scalable FPGA-based Accelerators on the Cloud

HiPEAC WRC
21 Jan 2019, Valencia

Dr. Chris Kachris
CEO, co-founder
www.inaccel.com

FPGAs in the news

May 29, 2018

Intel Delivers Xeon Scalable Processor 6138P with Arria 10 GX 1150 FPGA

Ratchets Up FPGAs in Data Center

by Kevin Morris

News & Analysis

Microsoft Eyes Expanding FPGA Role

Network chips not keeping pace

Intel, Alibaba Demo FPGAs in Cloud

March 10, 2017 by George Leopold

Nimbix Teams with Xilinx to Expand FPGA-Based Workload Acceleration in the Cloud

Xilinx Powers Huawei FPGA Accelerated Cloud Server

Baidu Deploys Xilinx FPGAs in New Public Cloud Acceleration Services

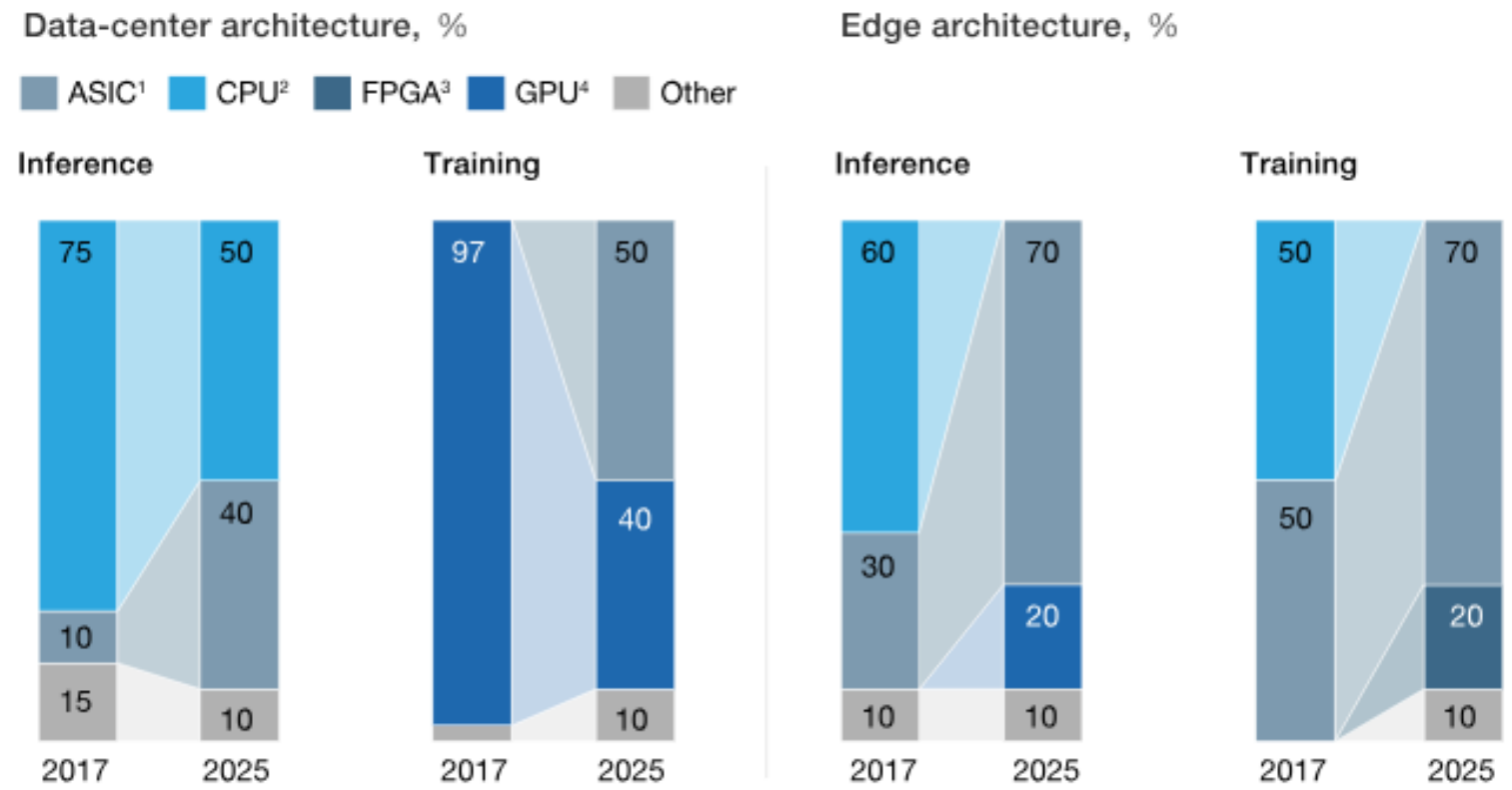
Market size for Reconfigurable Computing in the Cloud

> The **data center accelerator market** is expected to reach **USD 21.19 billion by 2023** from USD 2.84 billion by 2018, at a CAGR of **49.47%** from 2018 to 2023.

> The market for FPGA is expected to grow at **the highest CAGR during the forecast period** owing to the increasing adoption of FPGAs for the acceleration of enterprise workloads.

[Source: Data Center Accelerator Market by Processor Type (CPU, GPU, FPGA, ASIC)- Global Forecast to 2023, Research and Markets]

Artificial-intelligence hardware: New opportunities for semiconductor companies

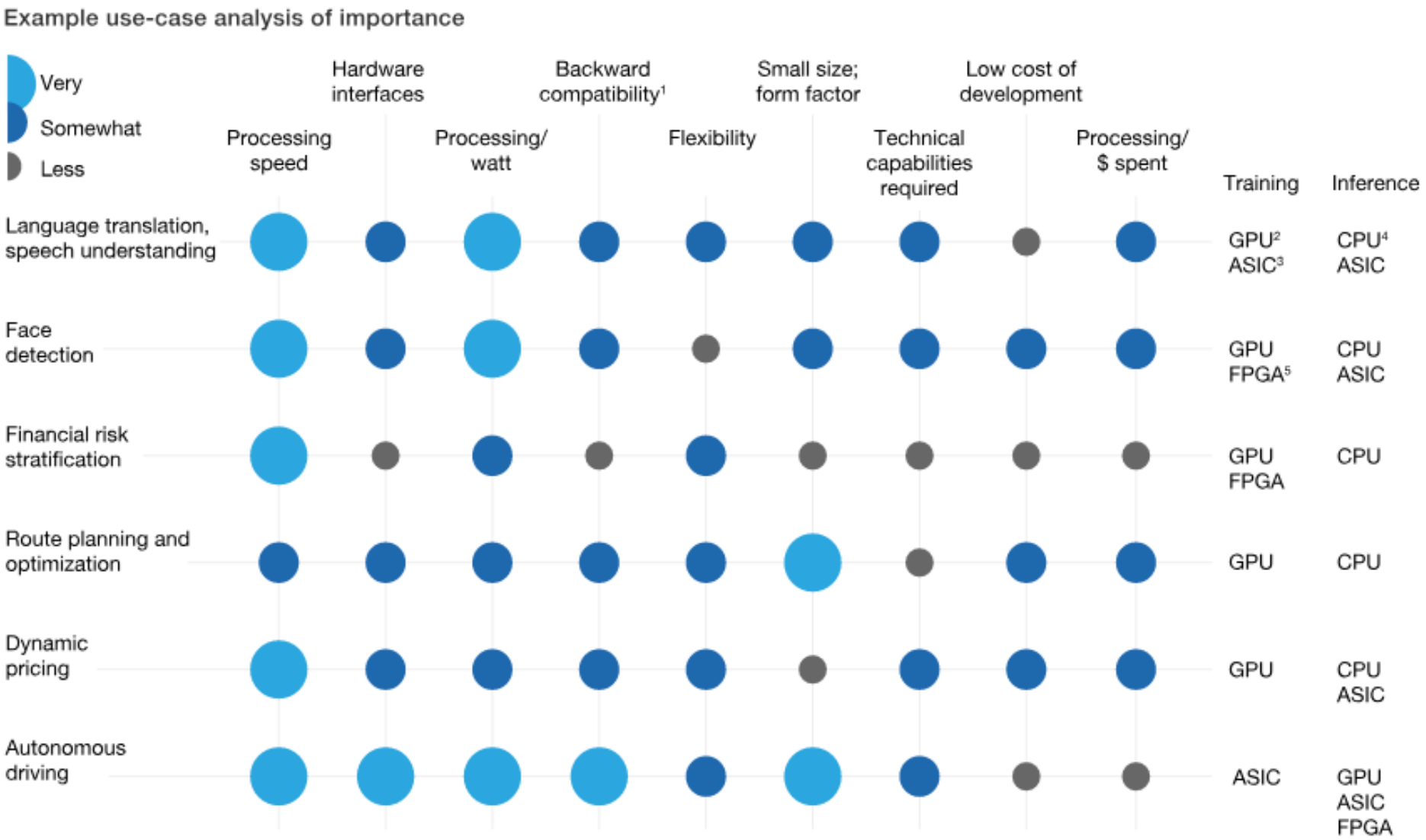


[1]Application-specific integrated circuit.
[2]Central processing unit.
[3]Field programmable gate array.
[4]Graphics-processing unit.

McKinsey&Company | **Source:** Expert interviews; McKinsey analysis

Souce: https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies.

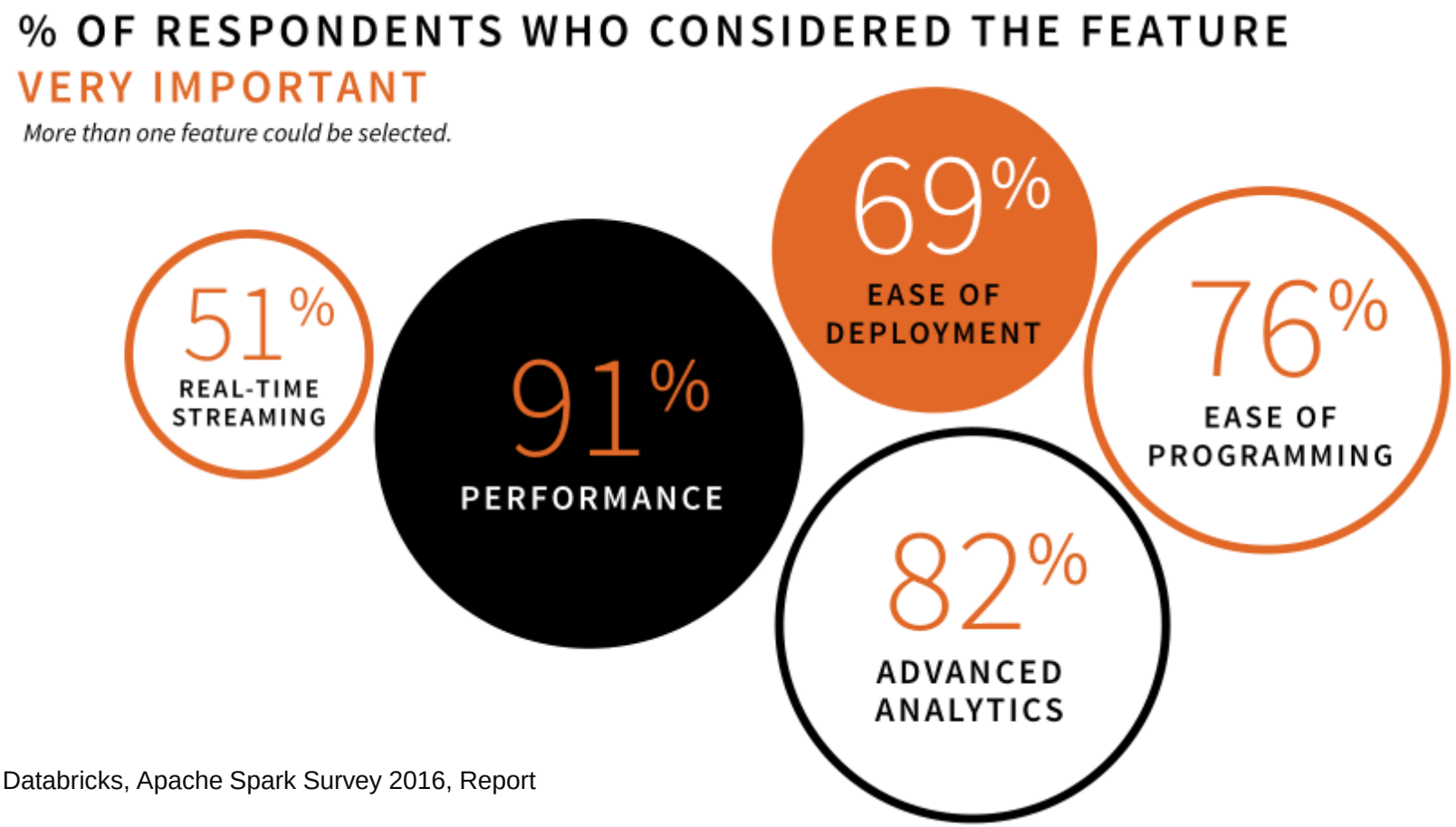
FPGA GPU ASIC



Souce: https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies.

Why acceleration

> **91% of Spark users for Big Data analytics care about Performance**



Source: Databricks, Apache Spark Survey 2016, Report

Acceleration for machine learning

inaccel offers Accelerators-as-a-Service for Apache Spark in the cloud (e.g. Amazon AWS f1) using FPGAs

ADVANCED ANALYTICS USERS (MLLIB)
IN PRODUCTION

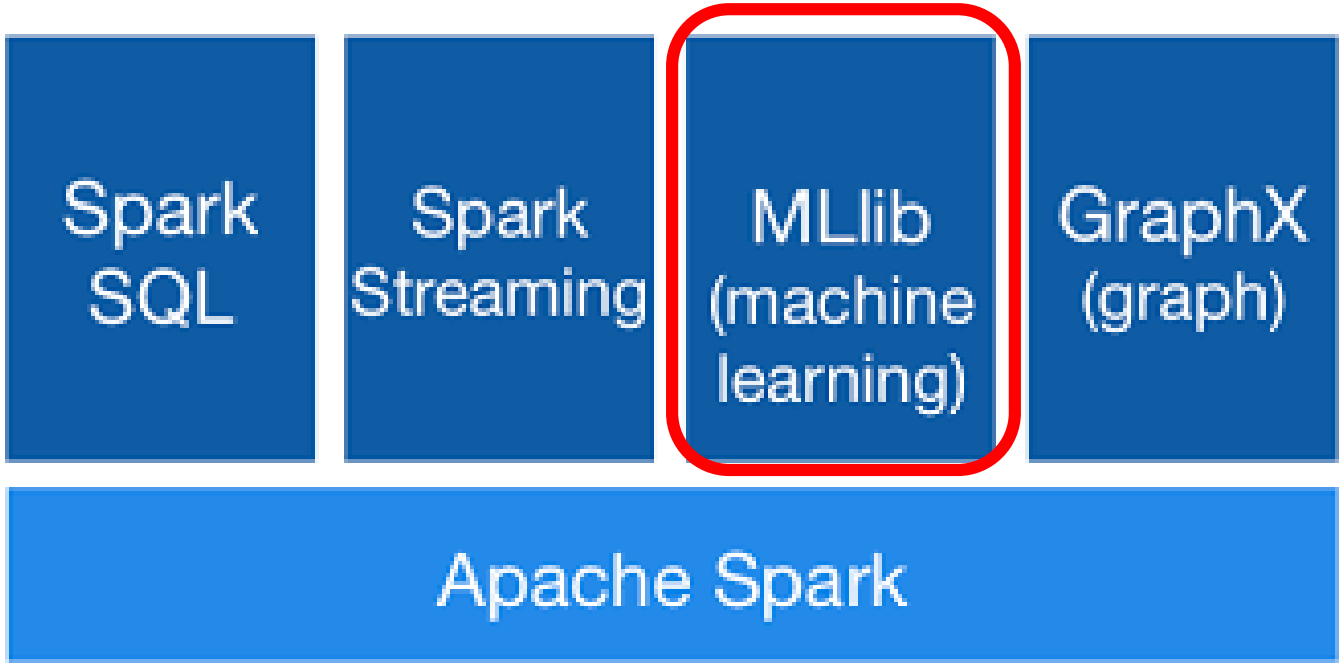
+38%

2015	2016
13%	18%
OF RESPONDENTS	OF RESPONDENTS

www.inaccel.com

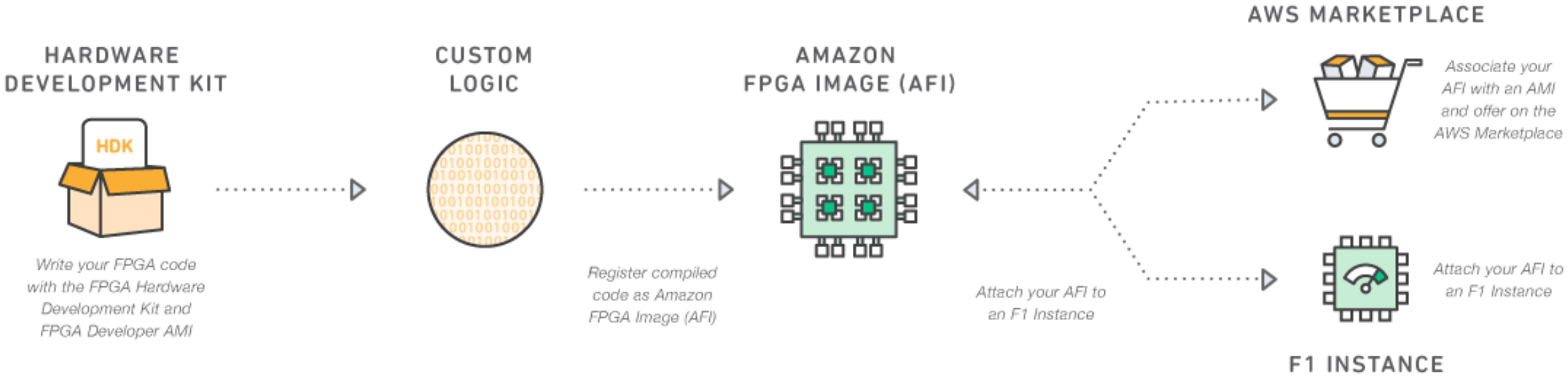
Apache Spark

- **Spark is the most widely used framework for Data Analytics**
- **Develop hardware components as IP cores for widely used applications**
  - Spark
    - Logistic regression
    - Recommendation
    - K-means
    - Linear regression
    - PageRank
    - Graph computing



FPGA in the cloud – the AWS model

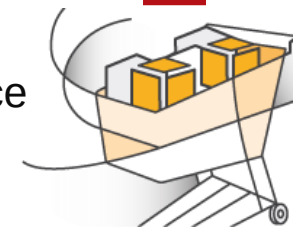
> Amazon EC F1's Xilinx FPGA

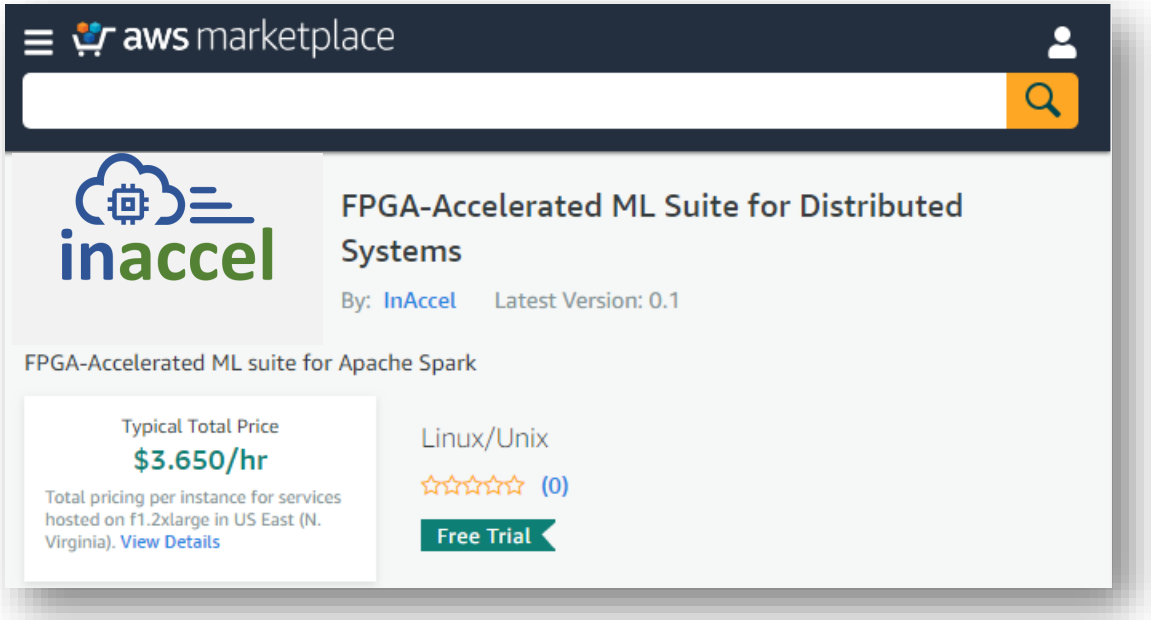


Cloud Marketplace: available now

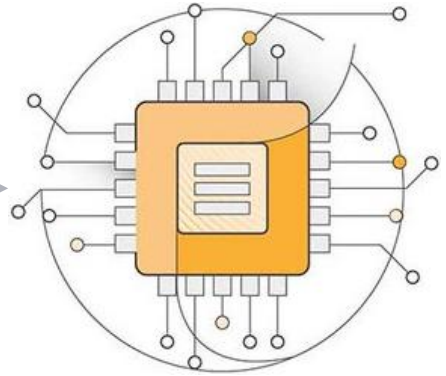
Customers

AWS Marketplace





InAccel Products



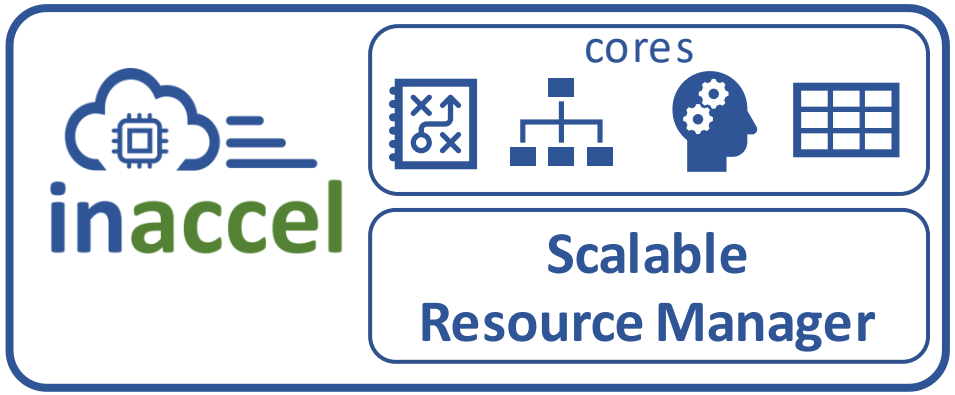
Amazon EC2 FPGA Deployment via Marketplace

Scalable to worldwide market

First to provide accelerators for Spark

www.inaccel.com

InAccel integrated framework



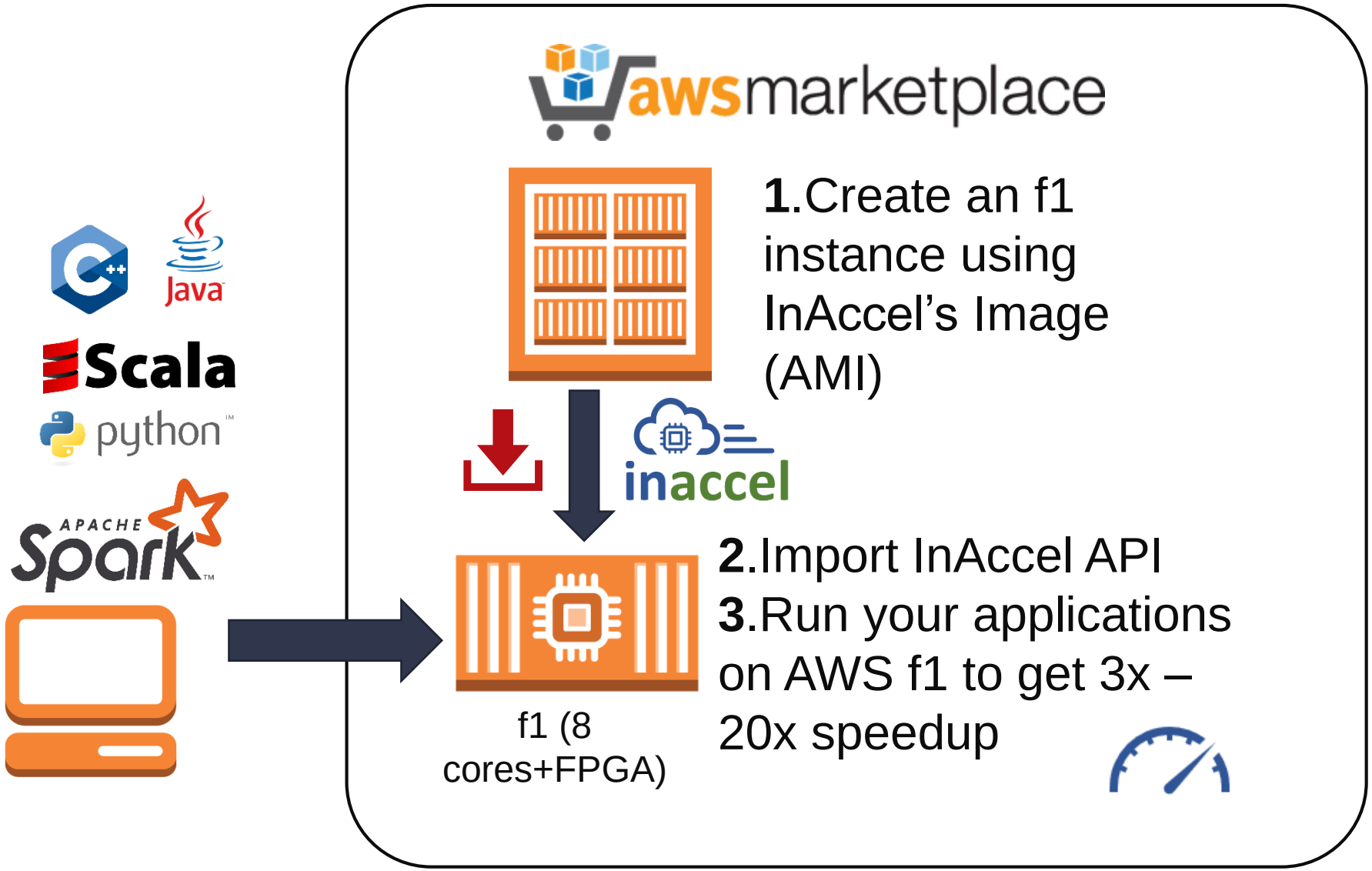
Supported APIs

Zero-code changes

- > **C/C++**
- > **Scala**
- > **Python**
- > **Java**
- > **Apache Spark**

www.inaccel.com

Accelerators for Spark ML in Amazon AWS in 3 steps



www.inaccel.com

IP cores available in Amazon AWS

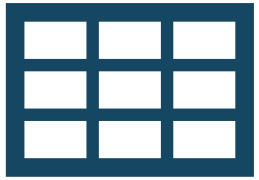
Logistic Regression

Gradient Descent IP block for faster training of machine learning algorithms.

K-mean clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.

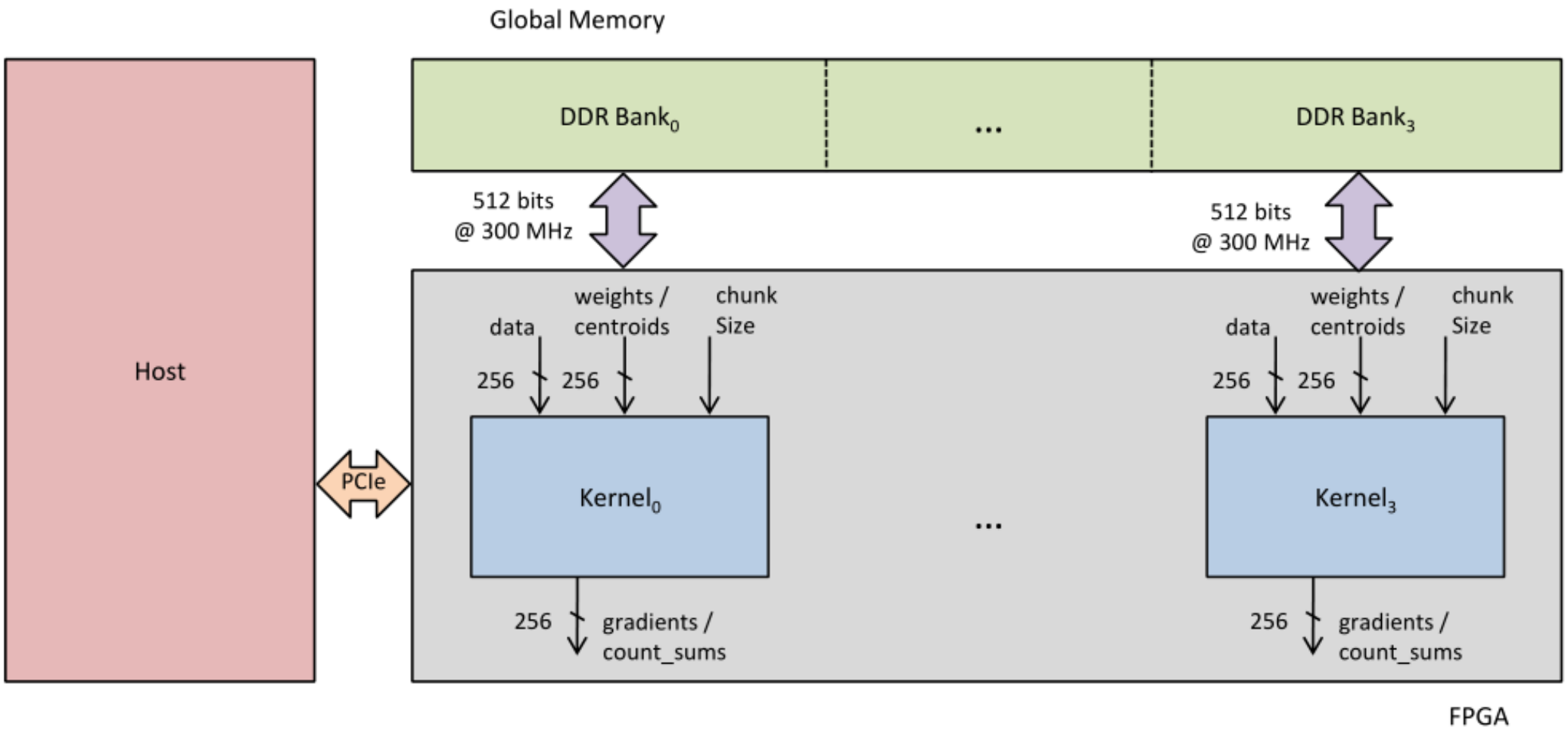
Recommendation Engines (ALS)



Alternative-Least-Square IP core for the acceleration of recommendation engines based on collaborative filtering.

Available in Amazon AWS marketplace for free trial: www.inaccel.com

Communication with Host in Amazon AWS



Accelerators for logistic regression/kmeans

www.inaccel.com

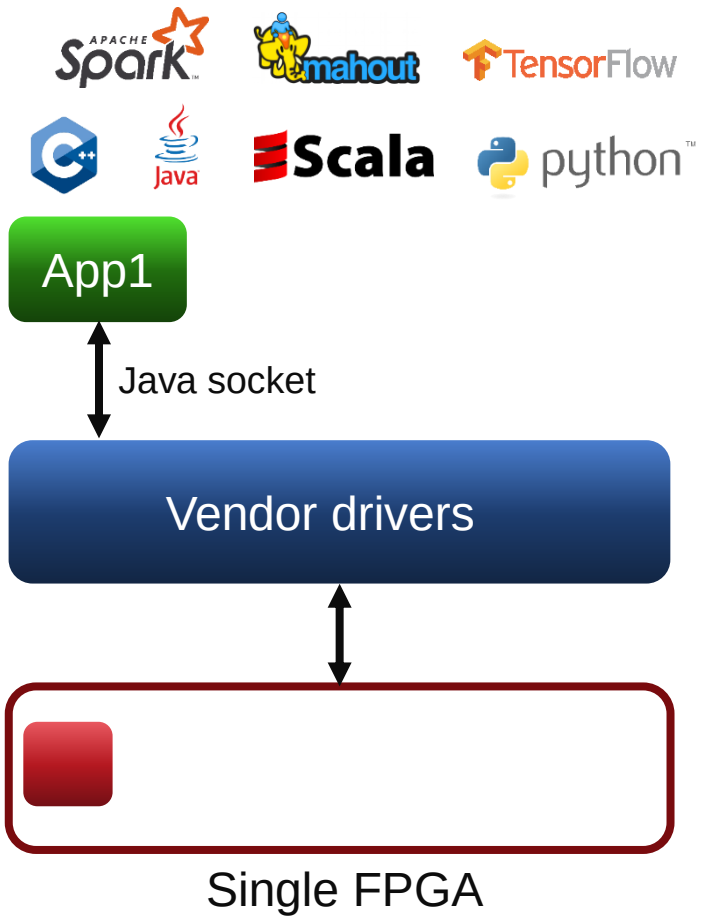
Current Framework for FPGAs on the cloud

Limitations

> Currently only **one application** can talk to each FPGA accelerator

> Every application can talk to a **single** FPGA.

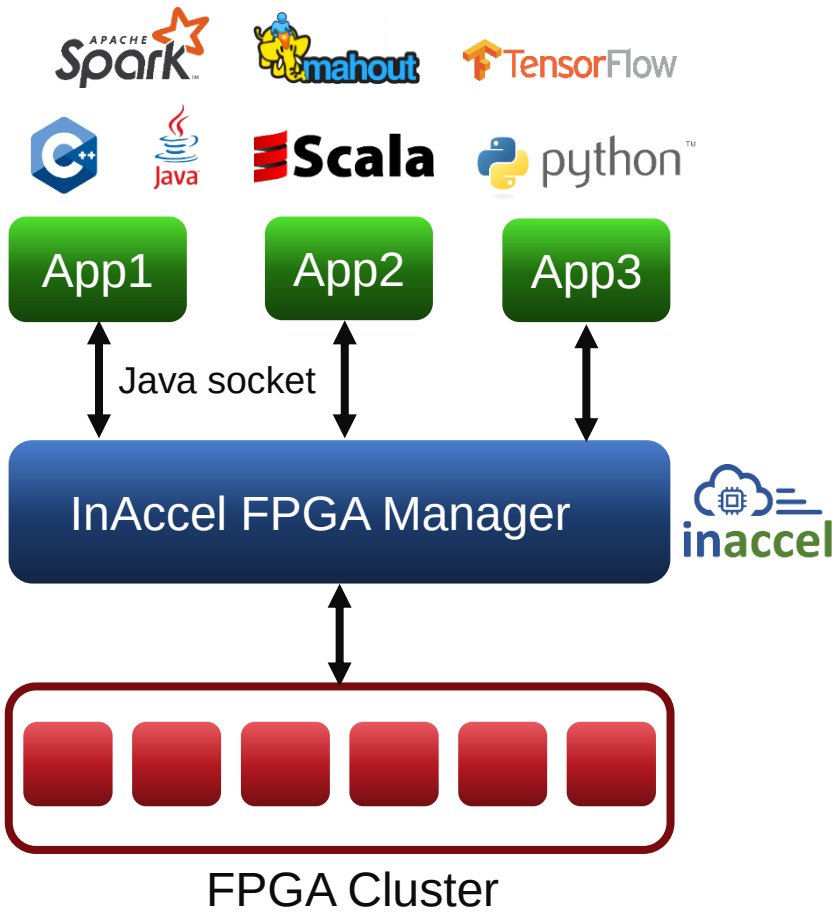
> Different architecture if you need to talk to multiple FPGAs



InAccel's Coral FPGA Manager

High-level abstraction layer to utilize and manage an FPGA cluster

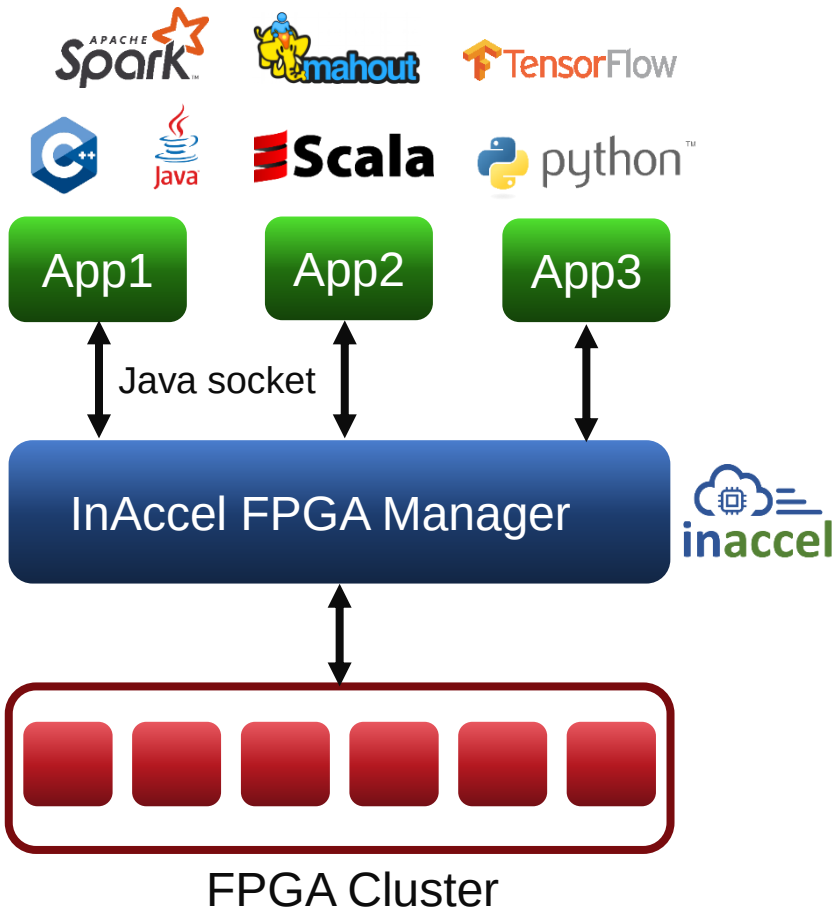
- **Resource Management**
  - Automatic configuration and management of the FPGA bitstreams and memory
- **Scheduling**
  - Automatic serialization and scheduling of the tasks send to the FPGA cluster
  - Scale-up to f1.x2, f1.x4, f1.x16 automatic
- **"Virtualization"**
  - Automatic serialization from multiple applications



FPGA Manager features

Ease of Use

> **Write applications quickly in Java, Scala and Python.**
> InAccel offers all the required high-level functions that make it easy to build and accelerate parallel apps. No need to modify your application to use an unfamiliar parallel programming language (like OpenCL)



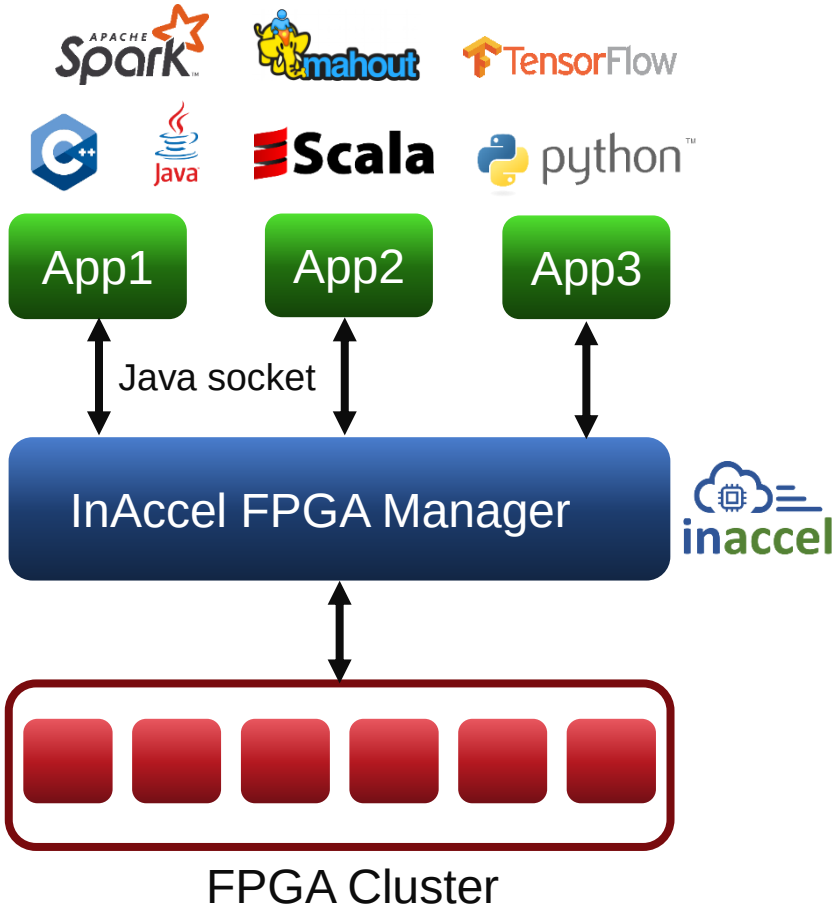
FPGA Manager features

Runs Anywhere

> Runs on any FPGA platform (**Xilinx, Intel**), giving you the freedom to take full advantage of on-premises, or public Cloud (**AWS, Alibaba, Nimbix, etc.**) infrastructure.

Alveo
U200

On-premise



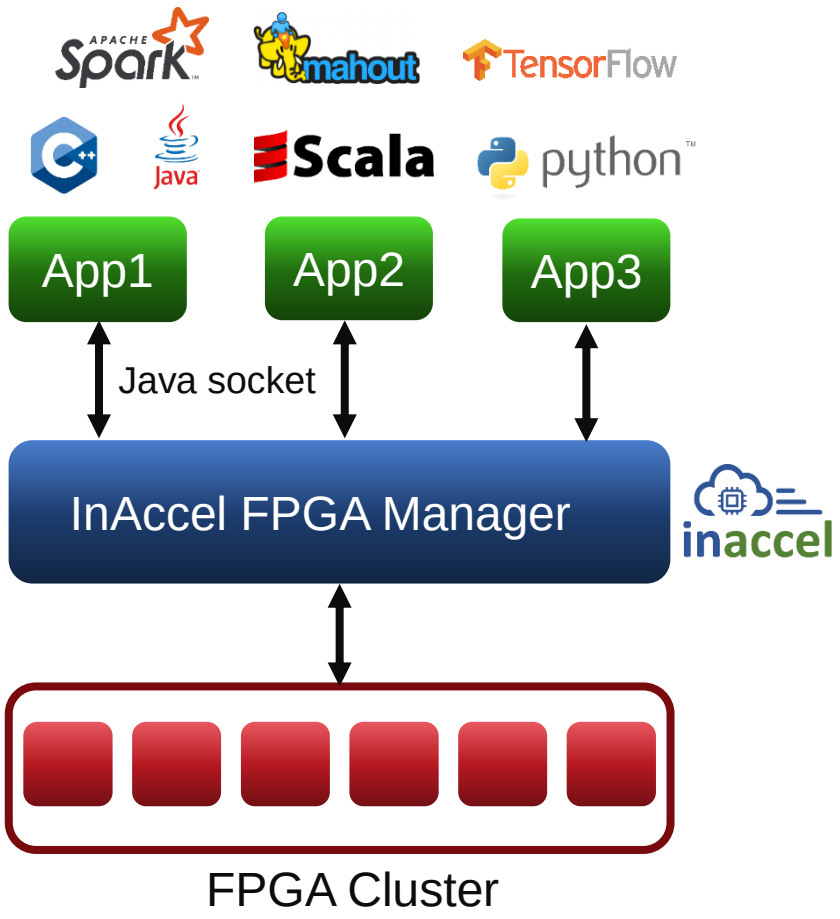
FPGA Manager features

Resource Management

> **Automatic resource configuration and task scheduling across entire FPGA clusters in private datacenters or public cloud environments.**
Coral examines the state of the FPGAs and implements load-balancing policies across them, efficiently taking care of all the required device configurations and memory transfers.

Privacy / Isolation

> **Coral allows the secure sharing of the hardware resources among different users and multiple processes or threads.**
First class isolation support for accelerator cores and FPGA memory.



InAccel's Run-time Engine

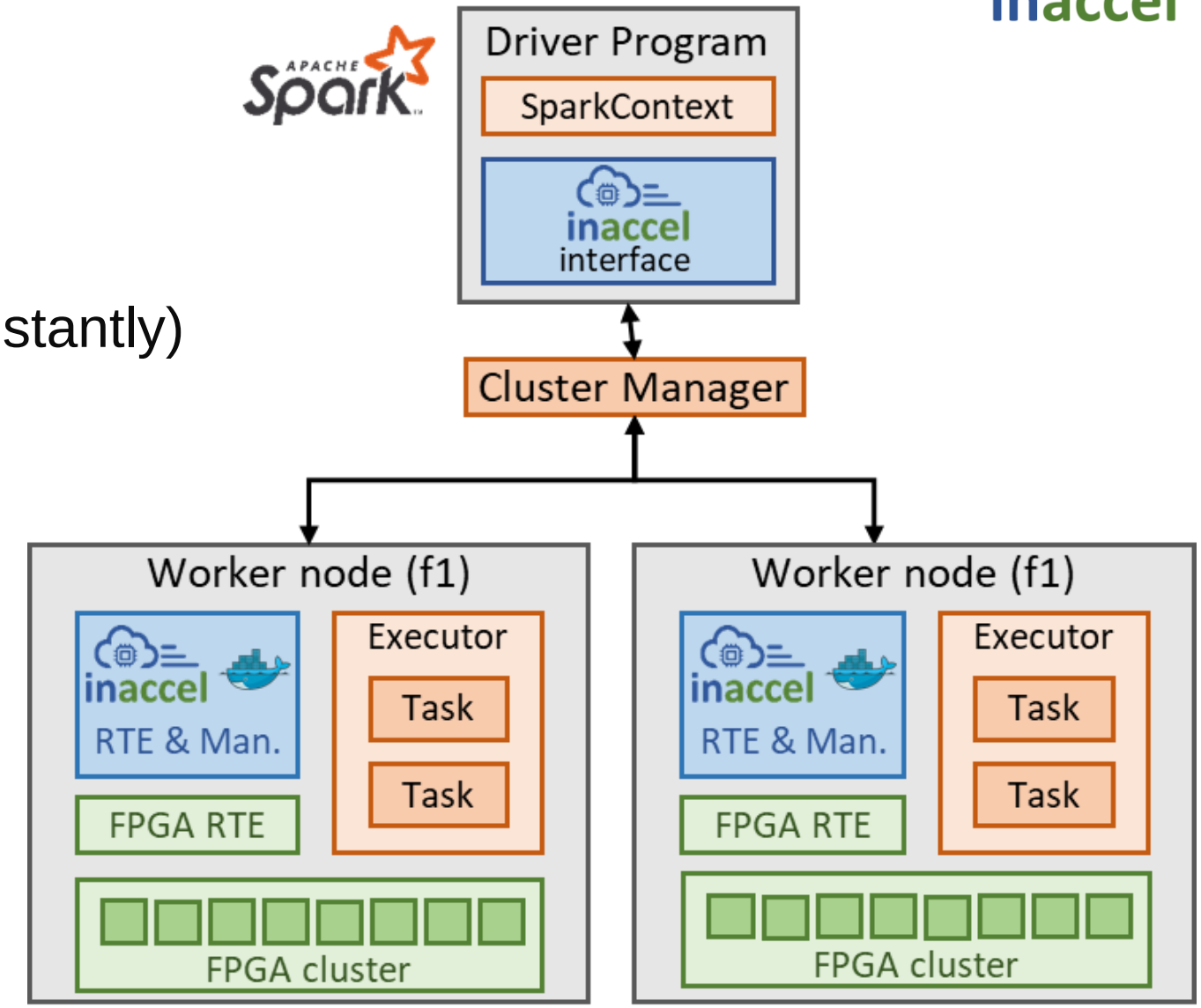
> **Runtime engine that allows**

>> **Scale Up** (1, 2, or 8 FPGAs instantly)

>> **Scale Out** (using Spark API)

>> Seamless integration

>> Docker-based deployment



FPGA Manager API

Memory Calls

> To make things easier we have incorporated a new **SharedMatrix** class that is basically backed up by a **Java ByteBuffer**.

Subclass	Used to store elements of type
SharedByteMatrix	byte
SharedDoubleMatrix	double
SharedFloatMatrix	float
SharedIntMatrx	int

Request Calls

> Request calls are responsible for sending new tasks to the FPGA manager. All the requests are static methods of **InAccel class**.

Request	Used to accelerate:
Gradients32	Logistic Regression
Centroids32	KMeans
Black-Scholes	Black-Scholes

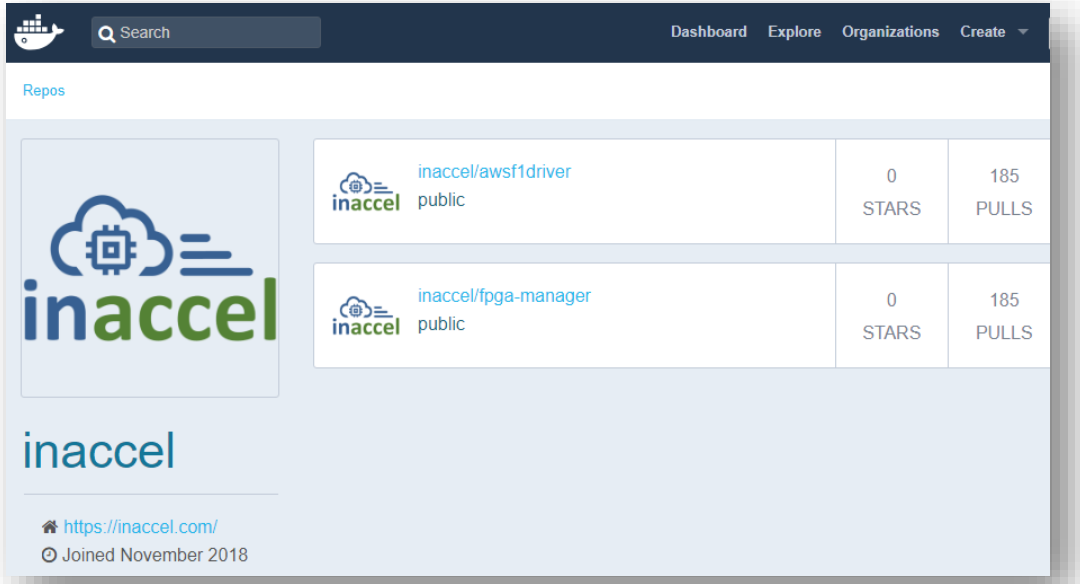
FPGA Manager deployment

> Easy deployment through dockers

> https://hub.docker.com/u/inaccel/

> Price for 3rd parties: $0.5/hour/node

> Free evaluation / limited features



inaccel

FPGA Manager

+

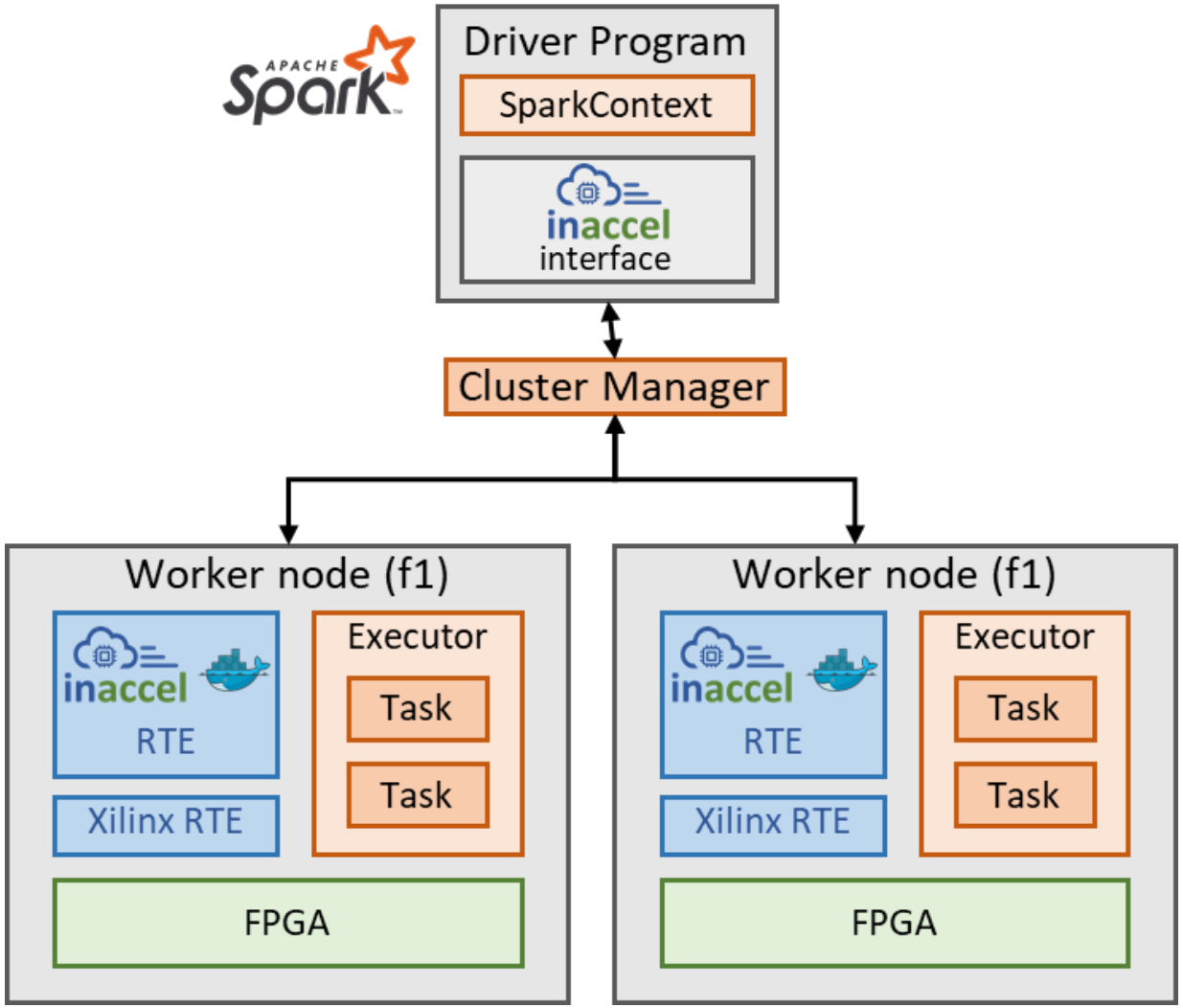
docker

=

- Easy deployment
- Easy scalability
- Easy integration

www.inaccel.com

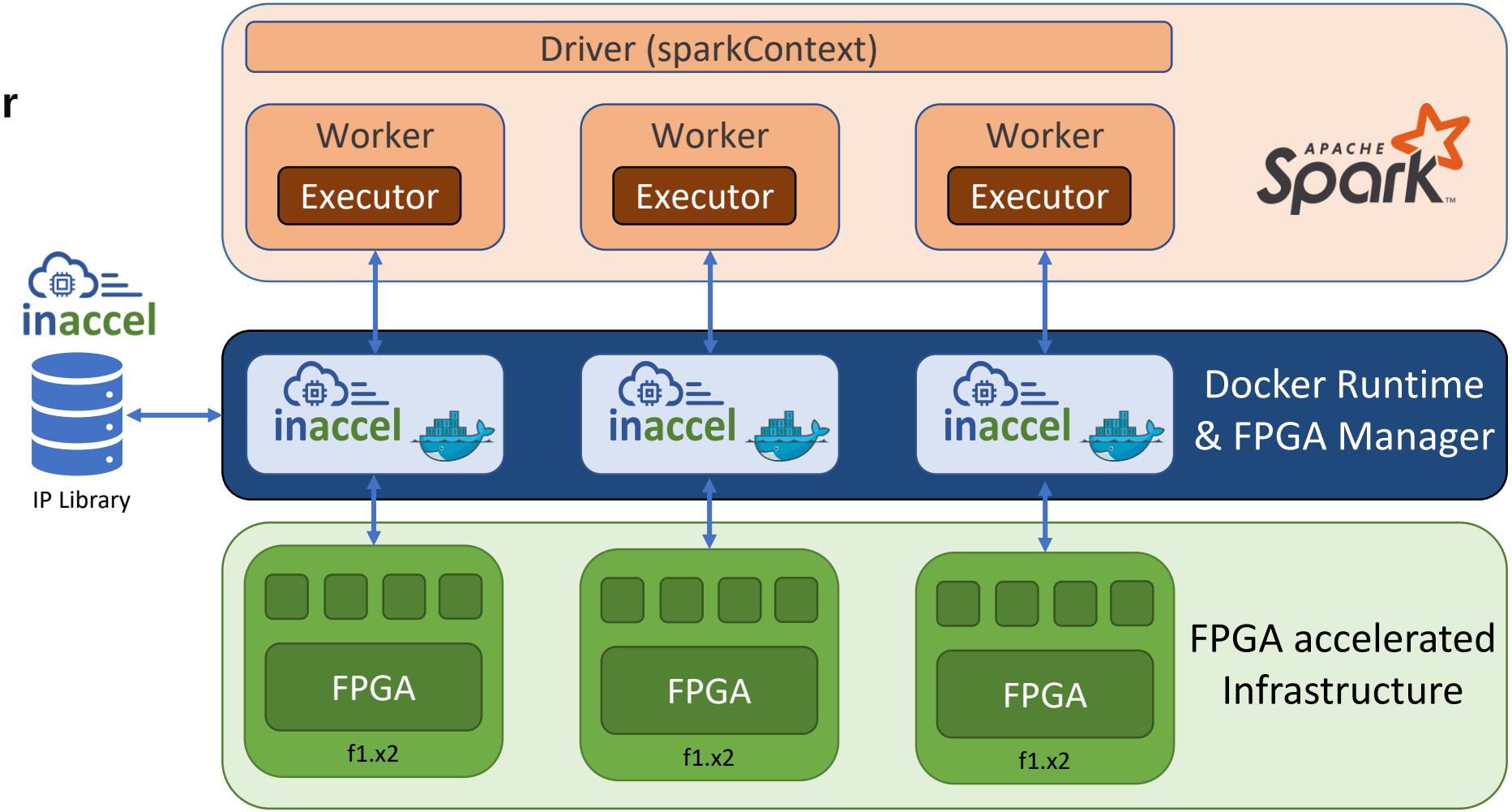
Docker-based implementation for easy integration



- Inaccel's FPGA manager docker container comprises both an FPGA manager to schedule, orchestrate, and monitor the execution of the accelerated applications but also the required FPGA runtime system.
- The dockerized runtime system detects the FPGA platform (aws F1) and manages the interaction/communication with the FPGA (i.e., loading the accelerator, transferring input data and results), making it transparent to the application.
- Docker swarm, Kubernetes, naïve execution

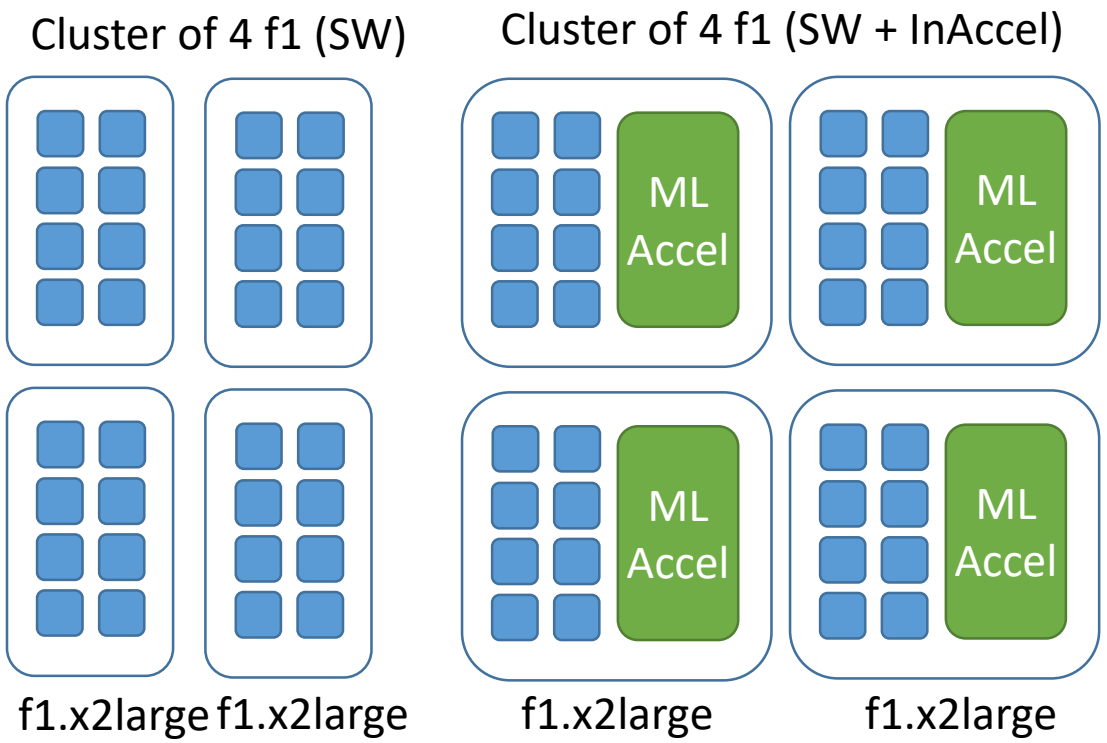
Cluster mode

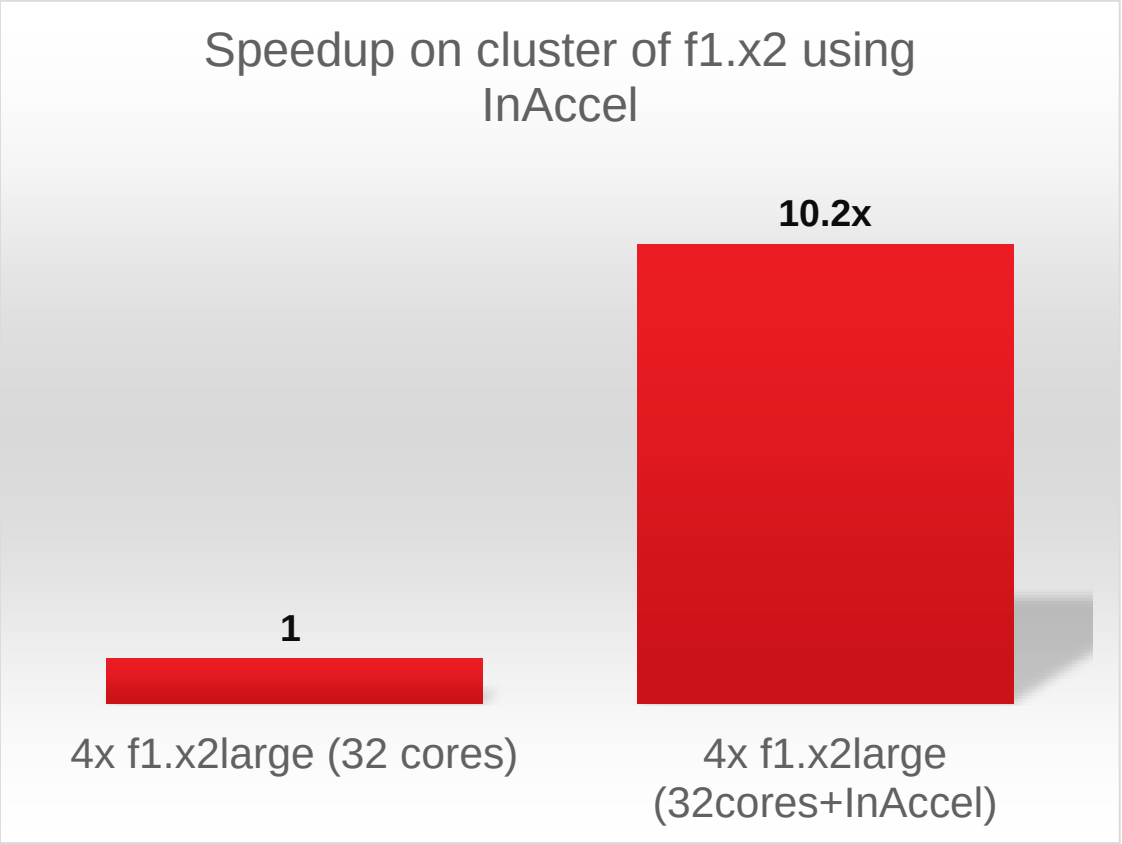
> **Cluster mode**



Speedup comparison

> **Up to 10x speedup compared to 32 cores based on f1.x2**

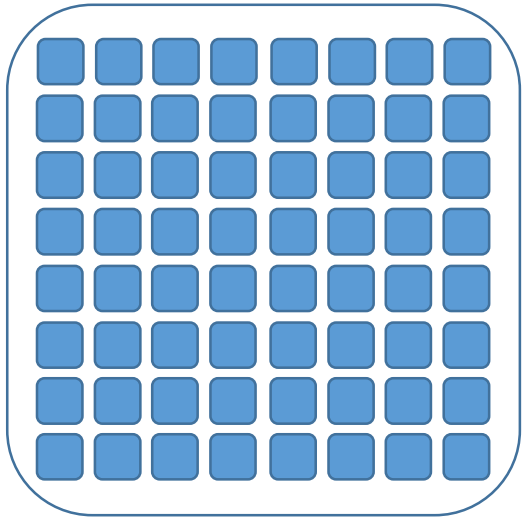




Speed up

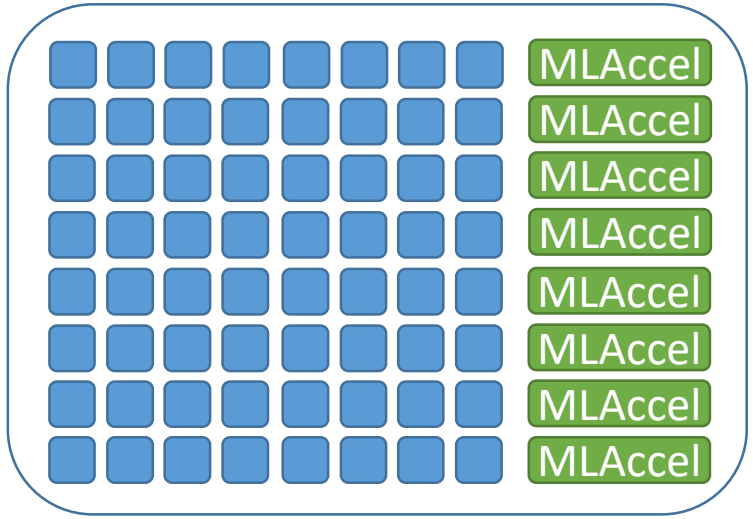
> **Up to 12x speedup compared to 64 cores on f1.x16**

f1.x16large (SW)

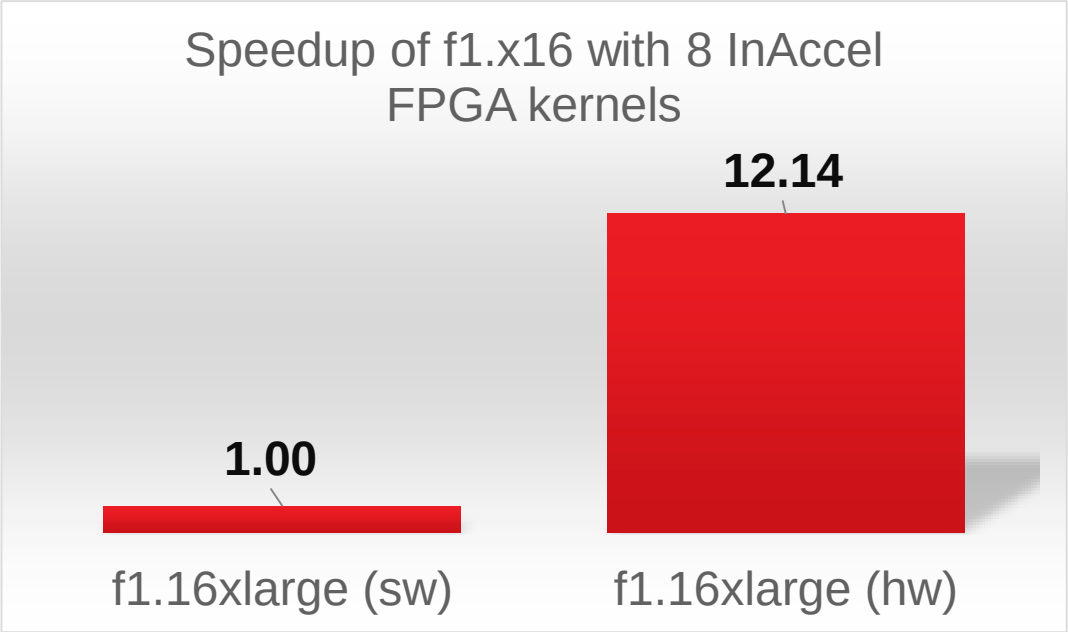


64 cores

f1.x16large (SW + 8 InAccel cores)



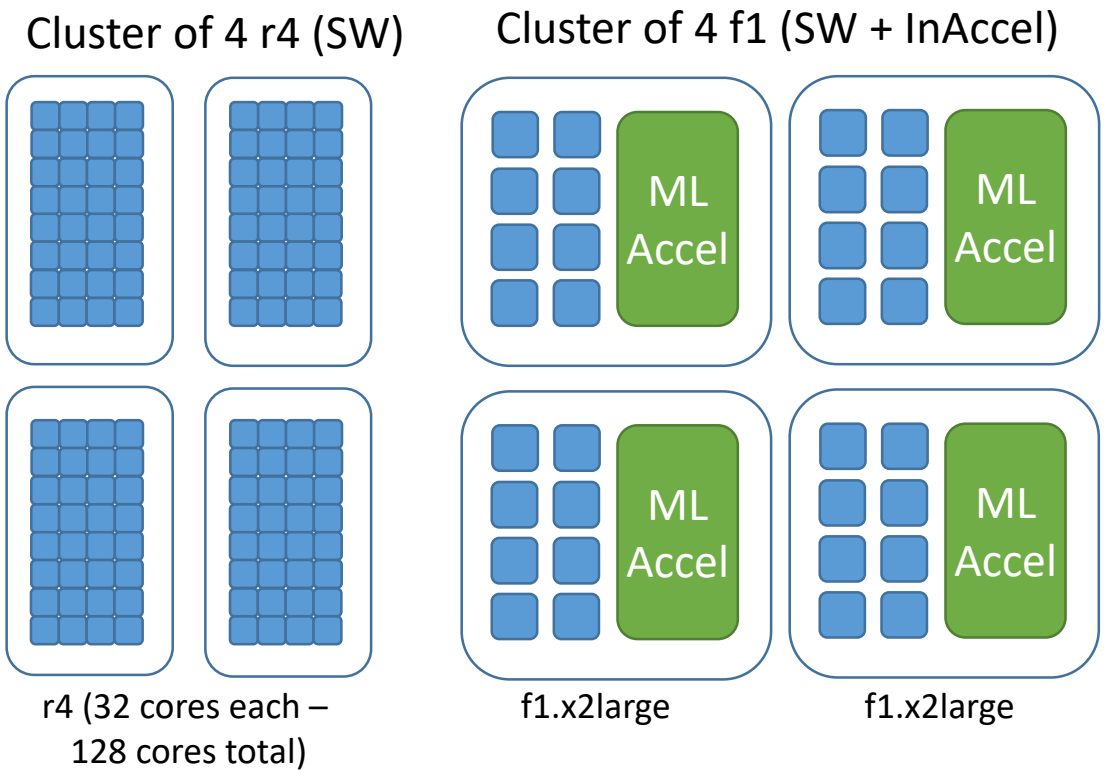
64 cores + 8 FPGAs with InAccel

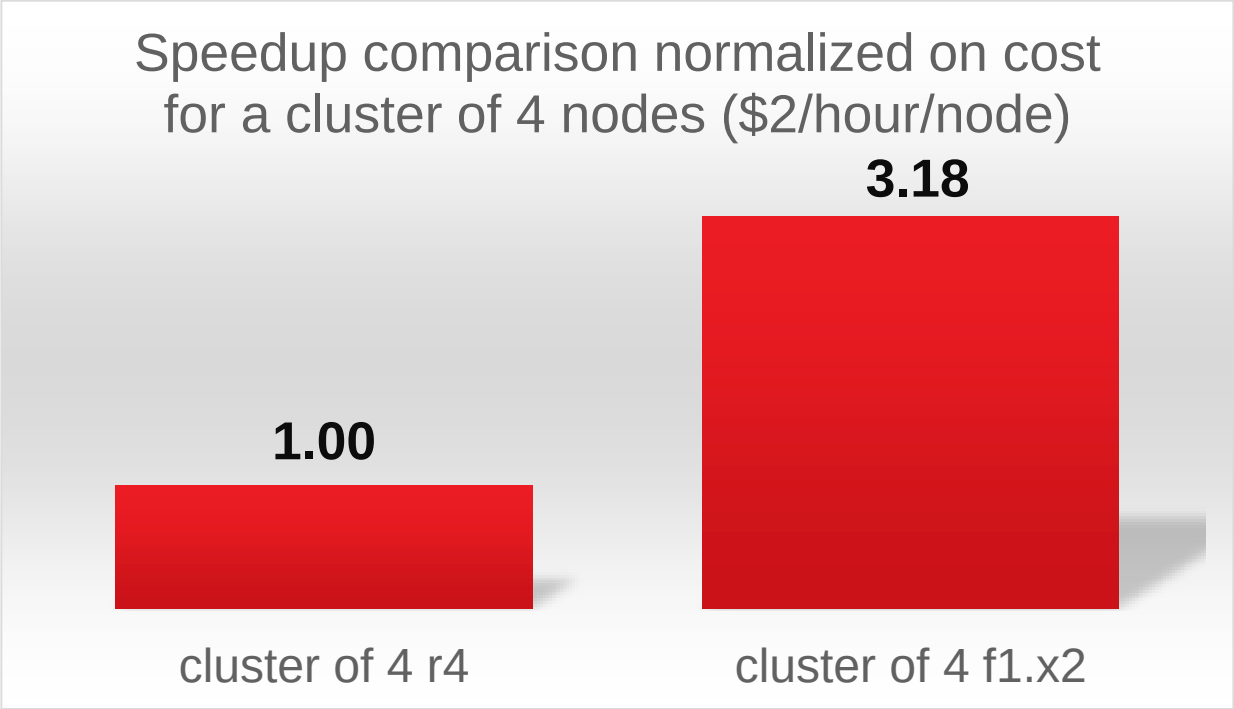


Speedup comparison

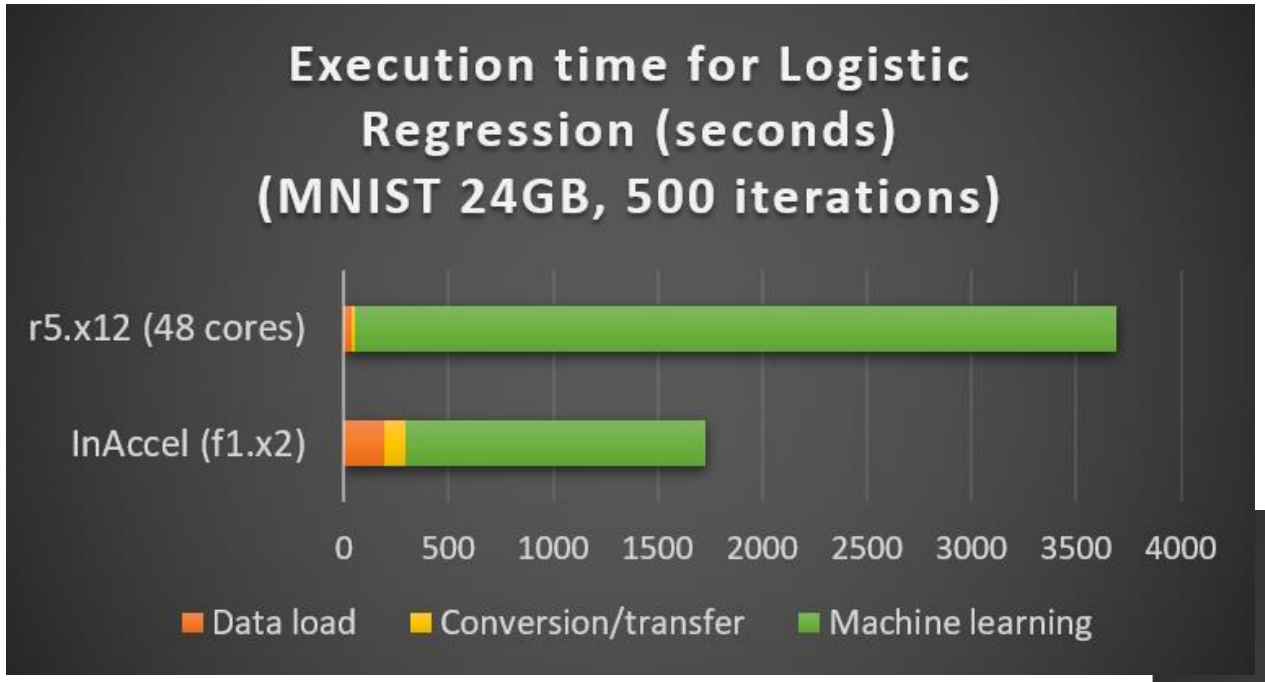
> **3x Speedup compared to r4**

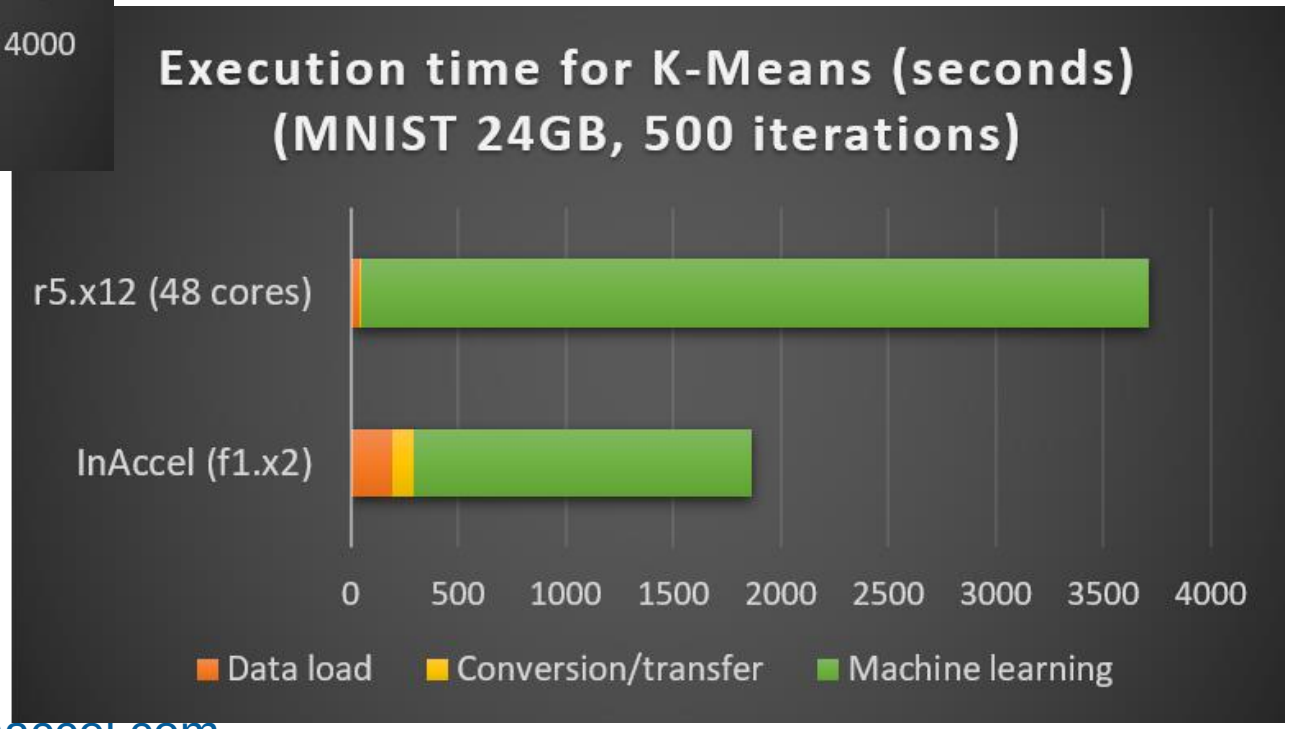
> **2x lower OpEx**





Performance evaluation





www.inaccel.com

Demo on Amazon AWS

Intel 36 cores Xeon on Amazon AWS
c4.8xlarge $1.592/hour

8 cores + inaccel
in Amazon AWS FPGA
f1.2xlarge $1.65/hour + inaccel

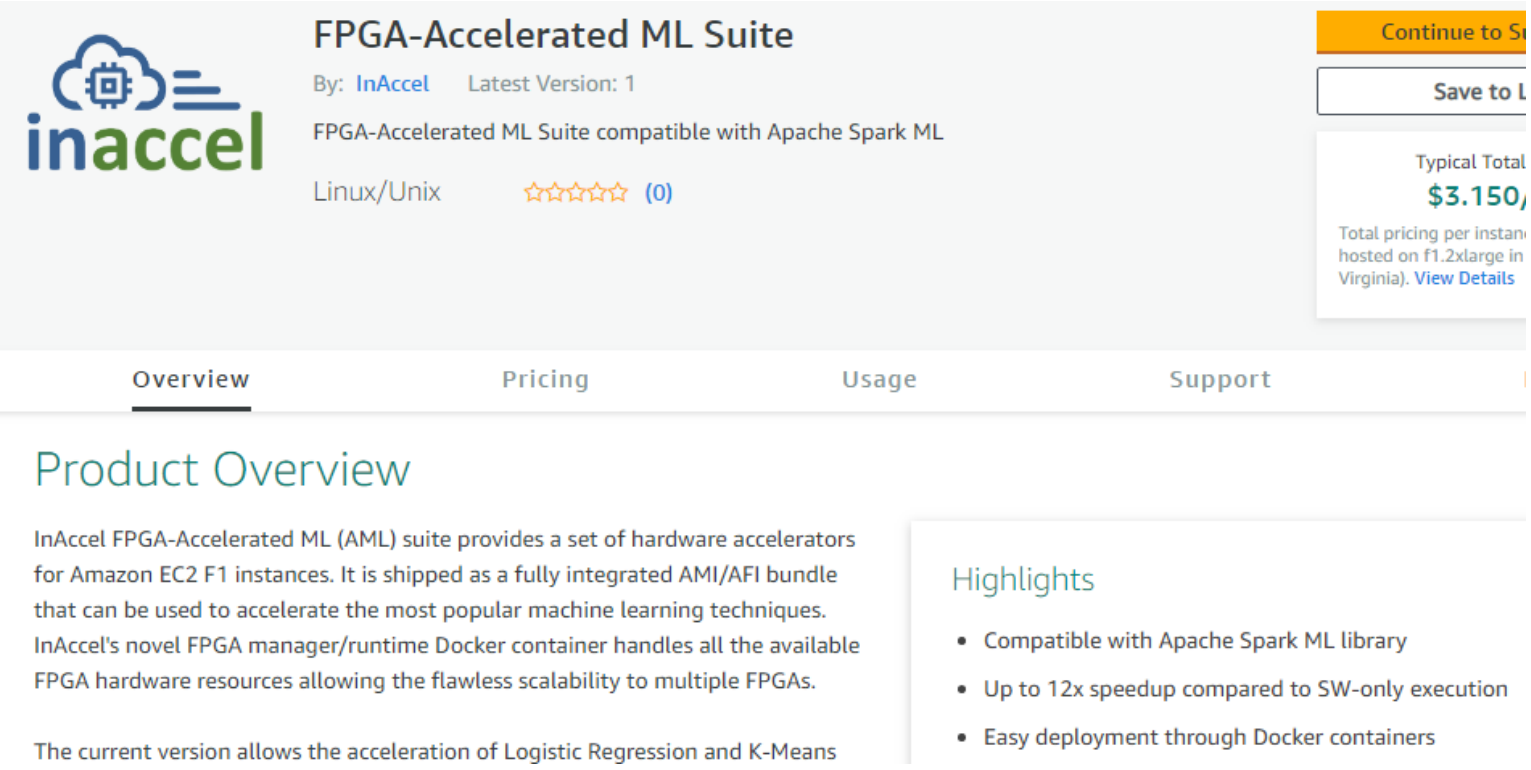
Note: 4x fast forward for both cases

www.inaccel.com

Cost reduction

> **Up to 3x lower cost to train your ML model**

Simple integration



> **InAccel OFF:** $ spark-submit [arguments]

> **InAccel ON:** $ spark-submit **--inaccel** [arguments]

InAccel unique Advantages

Compatible with Amazon AWS

All accelerators are compatible with the Amazon AWS F1 instances. AWS compatibility allows easy and fast deployment of the accelerators and seamless integration with your current AWS applications.

Seamless integration with your code

InAccel provides all the required APIs for the seamless integration of the accelerators without any modifications on your original code.

Acceleration of your code

Accelerators from InAccel provide up to 2x-10x speedup compared to contemporary processors in typical servers.

XILINX ALLIANCE PROGRAM MEMBER

aws partner network

awsmarketplace

www.inaccel.com

inaccel

helps companies speedup their applications

by providing ready-to-use accelerators-as-a-service in the cloud

3x-10x Speedup

2x Lower Cost

Zero code changes

www.inaccel.com

www.inaccel.com

chris@inaccel.com

#inaccel